融合角色、结构和语义的口语对话预训练语言模型

黄　健†，李　锋

(上海浦东发展银行股份有限公司 创新实验室，上海 200001)

摘　要：口语语言理解是任务式对话系统的重要组件，预训练语言模型在口语语言理解中取得了重要突破。然而，目前这些预训练语言模型，大多是基于大规模书面文本语料。考虑到口语与书面语在结构、使用条件和表达方式上的明显差异，构建了大规模、双角色、多轮次、口语对话语料，并提出融合角色、结构和语义的四个自监督预训练任务：全词掩码，角色预测，话语内部反转预测和轮次间互换预测，通过多任务联合训练面向口语的预训练语言模型 SPD-BERT：SPoken Dialog-BERT。在金融领域智能客服场景的三个人工标注数据集：意图识别、实体识别和拼音纠错上进行详细的实验测试，实验结果表明该语言模型的有效性。

关键词：对话系统；口语语言理解；预训练语言模型；意图识别；实体识别

中图分类号：TP183　　doi: 10.19734/j.issn.1001-3695.2022.01.0029

Spd-bert: role, structure and semantic based pre-trained spoken dialog language model

Huang Jian†, Li Feng

(*Innovation Lab*, *Shanghai Pudong Development Bank Co. Ltd*, *Shanghai* 200001, *China*)

Abstract: Spoken language understanding (SLU) is an important component of dialog system. Recently, pre-trained language model has made breakthrough in various tasks of spoken language understanding. However, these language models are trained with large-scale written language, which are quite different from spoken language in structure, condition and expression pattern. This paper construct large-scale multi-turn bi-role spoken dialog corpus. Then four self-supervised pre-trained tasks are proposed: masked language model, role prediction, intra-query reverse prediction and inter-query exchange prediction. A bert-based spoken dialog language model (SPD-BERT) is pre-trained through multi-task learning. Finally, the model is tested with three typical tasks of intelligent customer service in finance domain. The experiment results demonstrates the effectiveness of out model.

Key words: dialog systems; spoken language understanding; pre-trained language model; intent detection; named entity recognition

0　引言

对话系统在自然语言处理应用中扮演着重要的作用，取得了许多成功案例，如：智能客服，智能外呼，智能助手等，并广泛应用于金融、通信、电子商务等领域。通常来说，对话系统包括四大模块：自然语言理解(NLU，natural language understanding)，对话状态追踪(DST，dialog state tracking)，对话管理(DM，dialog management)和自然语言生成(NLG，natural language generation)。其中，口语语言理解是任务式对话系统[1~3]的重要组件，目的是从用户询问语句中获取关键的语义信息，包括众多细分任务：意图识别，实体识别[4]，情绪识别，态度识别等。与此同时，随着预训练语言模型[5](PTM，pre-trained language model)的发展，基于 PTM 的识别模型在口语语言理解的任务上取得了显著的效果，极大地提高了对话系统的客户满意程度。

传统的对话系统仅允许客户通过文本方式表达需求，这极大地限制了使用效率。为了提升客户体验，这些对话系统逐渐支持客户通过语音方式输入询问语句。并且，随着语音识别(ASR，audio speech recognition)技术的发展和成熟，越来越多的客户倾向于使用语音作为主要输入方式。客户的语音经过 ASR 转译为文本，并传递给对话系统。通过语音输入的文本，通常是口语化文本。

根据语言学的研究，口语与书面语存在差异，口语是听和说的语言，所以要求快，讲求效率，用词范围相对较窄，句子比较短，结构比较简单，有重复、脱节、颠倒、停顿等现象，还会出现语气词(如：嗯，呃等)。书面语是写和看的语言，这可以给人足够的时间进行推敲和琢磨。因此，口语化的文本语料和书面语文本语料存在显著的差异，图 1 展示了典型人与人口语对话案例,其中左侧为原始对话，右侧为 SPD-BERT 模型的输入和输出。

然而，目前预训练语言模型大多是基于书面语文本语料(例如：wiki，新闻等)训练得到。目前取得明显效果的口语语言理解模型，大多是直接基于这些预训练语言模型。再者，使用不同范式的语料训练获得语言模型，将学习到不同的知识。如果基于大规模口语化文本语料，训练语言模型，将进一步提高口语语言理解任务的效果。并且，书面语语料大多是基于长文本，不涉及角色转换。对于对话系统，往往是短文本，并且至少涉及两个角色的转换，从而导致在表达内容上呈现跳跃性。

因此，本文以 BERT 为核心骨架，训练面向口语对话的语言模型：SPD-BERT，即 SPoken Dialog BERT。本文的贡献总结如下：

1)构建大规模、双角色、多轮次、口语化对话语料。收集大规模领域对话语料，对 ASR 转译后的口语化文本进行清

收稿日期：2022-01-05；**修回日期：**2022-03-18

作者简介：黄健(1986-)，男(通信作者)，上海人，上海浦东发展银行股份有限公司创新实验室智能对话方向负责人，博士，主要研究方向为智能对话、自然语言理解、文本纠错、基于知识图谱的问答和产品推荐(jan8611@163.com)；李锋(1980-)，男，上海人，上海浦东发展银行股份有限公司创新实验室 AI 技术方向负责人，高级工程师，博士，主要研究方向为数字人、智能对话、语音识别、图像识别.

洗、合并、拼音纠错等处理，构建首个面向金融领域的千万级口语对话语料库。

2)创新性地提出角色、结构和语义融合的预训练任务。包括 4 个预训练任务：全词掩码(WWM, Whole Word Masking)，角色预测(RP, Role Prediction)，话语内部反转预测(IQRP, Intra-Query Reverse Prediction)，轮次间互换预测(IQEP, Inter-Query Exchange Prediction)。突破 BERT 的两个预训练任务(掩码和预测下一个句子(NSP，Next Sentence Prediction))的限制，提高角色、结构和语义的交互能力。

3)训练口语对话语言模型。基于大规模口语对话语料，将 4 个预训练任务联合学习，获得预训练口语对话语言模型 SPD-BERT，是业界首个面向金融领域多轮对话理解的预训练语言模型。

4)下游任务实验对比。基于预训练语言模型，在 3 个下游任务上进行实验对比分析，实验结果显示了 SPD-BERT 模型在口语语言理解任务上取得显著效果，在拼音纠错任务上，句子级别的 F1 提升 1.8%，显示出与书面语模型的明显优势。

本文的结构如下，第二部分介绍相关工作，第三部分详细阐述模型结构、4 个自监督预训练任务、多任务联合学习，并详细讨论大规模口语对话训练语料生成、以及模型训练等。第四部分对比分析多个口语语言理解任务的实验效果。最后，对全文进行总结，并对未来工作提出展望。



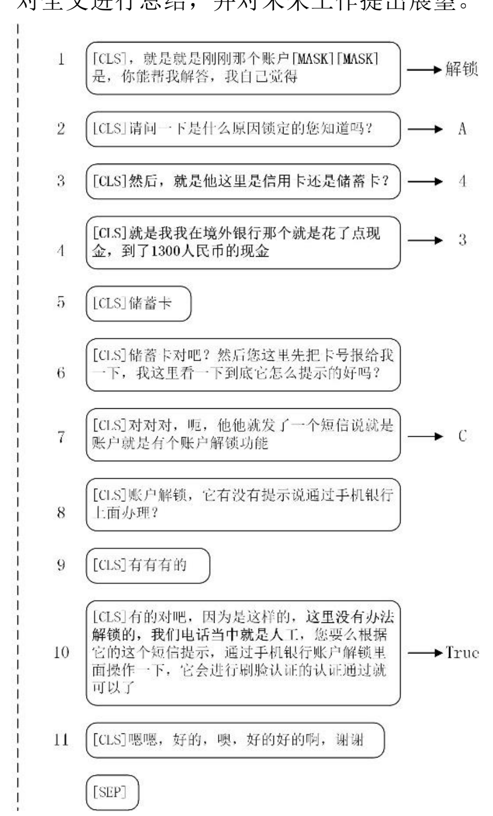


图 1　人与人口语对话案例

Fig. 1　Case of human to human conversation

1　相关工作

口语语言理解是对话系统的子任务，在任务式对话系统应用中起着非常重要的作用。近些年来，随着预训练语言模型的不断发展，基于预训练语言模型的口语语言理解模型效果得到持续的突破和提升。本部分将分别简要描述口语语言理解任务，以及对话预训练语言模型。

1.1　口语语言理解

口语语言理解通常包括若干子任务：意图识别，槽位填充，拼音纠错，实体识别等。通过意图识别，理解客户的意图，进入多轮对话流程；通过槽位填充，获取任务的关键、必要信息；通过情感识别，了解客户此时的满意程度；同时，为了减少因 ASR 转译错误导致的识别错误，通过拼音纠错，提升客户输入文本的质量。

传统情况下，将意图识别和槽位填充作为独立的两个任务来训练。文献[6]提出基于循环神经网络(RNN，LSTM 等)实现意图识别，结果表明，序列特征能够显著提升意图识别效果。文献[7]提出基于循环神经网络的编码器，使用句子级别的信息，提升槽位填充任务的效果。考虑到意图识别与槽位填充任务之间的内在关系，使用两个任务共享的知识，可以显著提升两个任务的效果，联合训练模型逐步得到发展。文献[8]提出槽位填充-意图识别双向交互网络模型来建立两个任务之间的直接关系，考虑到意图识别对槽位填充的影响，以及槽位填充对意图识别的影响，从而互相提高效果。文献[9]通过构建意图识别和槽位填充任务的双向联系，提出联合交互模块来实现两者的互相影响，该方法基于 transformer 特征提取器，并设计了精巧的交互注意力层，取得了显著的效果。随着预训练语言模型在各种自然语言处理任务上的突破，口语语言理解的研究也逐步探索基于 BERT[5]的语义理解模型。文献[10]提出基于 BERT 的意图识别和槽位填充联合训练模型。文献[11]提出基于 BERT 的多语言文本分类和序列标注联合框架。基于 BERT 的意图识别和槽位填充模型，取得了较为显著的效果提升。近期，文献[12]对口语语言理解进行了详实的综述，这里不再赘述。

口语语言理解处理的文本大多是经过 ASR 转译后的口语化文本，而中文存在大量同音字，因此，拼音纠错对口语语言理解的整体效果起着非常重要的作用。早期的研究主要采取流水线方式：错误识别，候选生成和结果选择。文献[13]使用基于字符的 N 元语言模型来检测潜在错误拼写字符集，并基于拼写和拼音相似度生成候选集，最后根据语言模型概率选择最佳候选。文献[14]使用掩码语言模型作为去噪自编码器来生成候选集，并提出置信度和似性解码器来过滤候选集。文献[15]提出基于图卷积网络的拼写纠错模型，基于同音字图网络和同形字网络，可以学习到每个字的语义表示，并

作为 BERT 的输入向量，从而学习到更丰富的句子语义表示。文献[16]从字符、位置、拼音、笔画四个维度来表示每一个字符，并通过困惑集来生成掩码训练数据，从而得到包含拼写错误知识的预训练语言模型。

1.2　对话预训练语言模型

自然语言表示学习从早期的基于统计的 N 元模型，到分布式表示[17,18]。这些属于静态词向量，即用一个固定的向量来表示某个词。然而，由于语言的灵活性和高效性，自然语言中存在大量的同义词。为了解决一词多义的问题，文献[19]提出基于双向训练神经网络的考虑上下文信息的词向量表示方法，缓解了多义词的表示问题。随着更强大的特征提取器 Transformer[19]的提出，自回归语言模型 GPT[20]和自编码语言模型 BERT[5]不断刷新各种自然语言任务的最优效果。

值得注意的是，这些预训练语言模型的训练数据，大多是大规模书籍语料和维基百科等文档型书面文本，而非口语化文本。近期，针对对话系统中口语化文本的特点，许多研究者提出面向对话口语化的预训练语言模型。文献[21]认为任务式对话系统的语言模式与通用文本存在显著的差异，整合大规模人人、多轮、任务型对话数据集，将用户和系统标识融入到掩码语言模型中，与 BERT 对比发现，在 4 个下游任务中取得显著效果。文献[22]引入语音和文本，提出跨模态掩码语言建模任务和跨模态条件语言建模任务，来支持端到端口语语言理解。文献[23]提出多角色对话理解预训练语言模型，通过设计若干自监督任务，尝试从对话中学习“谁对谁说了什么”，从而提高对话理解过程。文献[24]基于层次化循环编码器-解码器，来编码上下文信息，从而能够生成语义更加流畅的回答。本文与上述研究成果的区别在于：预训练任务的类型。

2　SPD-BERT 模型

本文提出的面向口语对话预训练语言模型 SPD-BERT，需要理解每一轮次的角色，以及该角色的话语语义。因此，其输入可以表示为：$d_k=\{(s_i,u_i)\}_{i=1}^{m}$，其中，$d_k$ 表示某次完整对话，m 表示对话的总轮次，$i\in[1,m]$ 表示第 i 轮次，s_i 表示第 i 轮次的角色，u_i 表示第 i 轮次的角色所说的话语。u_i 可以进一步表示为：$\{u_j^i\}_{j=1}^{n_i}$，其中，n_i 是第 i 轮的话语长度，u_j^i 表示第 i 轮话语中的第 j 个字符，$j\in[1,n_i]$。本模型的目的是，给定任意一通对话，为每一角色说的话语，结合上下文，生成其嵌入向量。值得注意的是，该嵌入向量，不仅包含话语的上下文语义和结构信息，还包括对应角色的信息。因此，该嵌入向量能够通过微调有效地应用到不同的下游任务。

2.1　模型概览

模型的输入表示和模型的整体结构，如图 2 所示。输入表示包括三个部分：token 编码、片段编码和位置编码。Token 编码和位置编码采取 BERT[5]模型中的编码方式。与传统片段编码不同的是，本文的模型是面向多轮次、双角色对话场景，因此，这里的片段数量与总的对话轮次成正相关。

将 token 编码与片段编码、位置编码相加，作为 transformer 的输入表示。其中，token 编码为 $\boldsymbol{e}_{ij}^{t}\in\boldsymbol{R}^{d}$，对应的字符嵌入表为 $\boldsymbol{E}^{t}\in\boldsymbol{R}^{V\times d}$，其中，V 表示词表大小；片段编码为 $\boldsymbol{e}_{ij}^{s}\in\boldsymbol{R}^{d}$，对应的片段嵌入表为 $\boldsymbol{E}^{s}\in\boldsymbol{R}^{S\times d}$，其中，S 表示最大片段数量(即最大对话轮次数量)；位置编码为 $\boldsymbol{e}_{ij}^{p}\in\boldsymbol{R}^{d}$，对应的位置嵌入表为 $\boldsymbol{E}^{p}\in\boldsymbol{R}^{N\times d}$，其中，N 表示整个对话的序列长度，即 $\mathrm{N}=\sum_{i=1}^{m}n_i$。这里，d 设置为 768。因此，transformer 的输入表示为

$$\boldsymbol{e}_{ij}=\boldsymbol{e}_{ij}^{t}+\boldsymbol{e}_{ij}^{s}+\boldsymbol{e}_{ij}^{p} \tag{1}$$

其中，$\boldsymbol{e}_{ij}\in\boldsymbol{R}^{d}$。经过 transformer 的强大特征提取能力，输出每个位置对应的嵌入向量：

$$\boldsymbol{E}_{ij}=\boldsymbol{transformer}(\boldsymbol{e}_{ij}) \tag{2}$$

这里，$\boldsymbol{E}_{ij}\in\boldsymbol{R}^{d}$，表示序列每个字符的输出向量。利用每个片段的第一个嵌入向量(即 $\boldsymbol{E}_{CLS_i}=\boldsymbol{E}_{i1}$)，经过非线性分类器，可以识别该片段的角色、轮次、是否存在内部反转等。利用每个片段的其他位置的嵌入向量，可以判断是否存在掩码，以及掩码对应的实际文本。

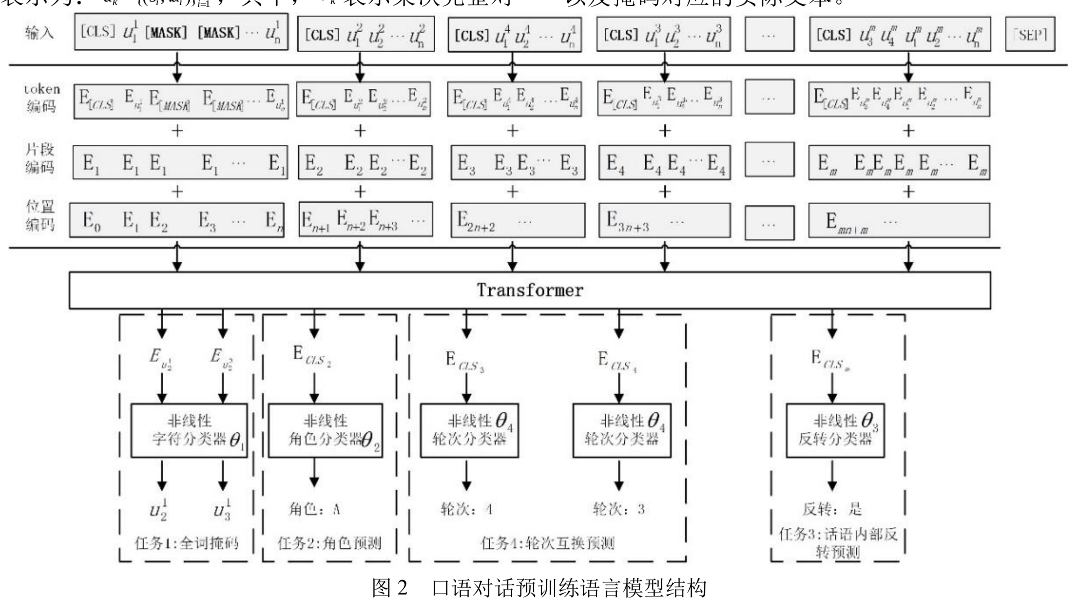


图 2　口语对话预训练语言模型结构

Fig. 2　Model architecture of SPD-BERT

2.2　预训练任务

为了充分挖掘大规模多轮、双角色对话中蕴涵的知识，基于传统预训练任务的两大核心：掩码和顺序，本文提出角色和语义融合的四项预训练任务。将这四项自监督预训练任务联合训练，在进行语义建模的同时，充分考虑话语角色和对话结构上下文。注意的是，这里的自监督是指，不需要对大规模语料进行人工标注，而是直接利用语料本身的标签，例如：某 query 是哪个角色说的，可以构建角色标签；某角色先说了某 query，下一轮对话中说了另外一个 query。可以构建 query 内部的顺序标签，以及轮次之间的顺序标签。

2.2.1 全词掩码 WWM

常规的掩码语言模型，是 15%概率选择输入序列的字符，进行掩码。然后，针对这些字符，以 80%概率实际进行掩码，10%概率随机替换，10%概率保持不变。为了提高模型的语义

学习能力，文献[25]提出了全词掩码(WWM, Whole Word Masking)，基于预设词典，将连续的若干字符，同时掩码。这里，本文采取的是全词掩码。例如，对于角色 s_i 的话语 $u_i=\{u_1^i,u_2^i,u_3^i,\ldots,u_n^i\}$，如果选择其中的第 2、3 个字符需要掩码，即得到 $u_i=\{u_1^i,[MASK],[MASK],\ldots,u_n^i\}$。图 1 案例中的第一轮对话中的"解锁"，表示了全词掩码。利用每个位置输出的嵌入向量，通过非线性字符分类器，预测字符，并与实际字符进行比较。

$$p_{ij}^1=\mathrm{softmax}(E_{ij}\cdot E^t.\mathrm{T}+b_1) \tag{3}$$

其中，$p_{ij}^1\in \boldsymbol{R}^V$ 表示对掩码后的 u_j^i 的预测值，b_1 为非线性分类器的偏置参数。值得注意的是，这里共享了字符嵌入表 E^t，只是在计算过程中进行了转置。模型的编码器参数记为 θ，非线性字符分类器参数记为 θ_1，输入序列掩码的字符数量为 M。全词掩码预训练任务的损失函数可以表示为

$$L_1(\theta,\theta_1)=-\sum_{k=1}^{M}logp_{ij}^1(m=m_k\,|\,\theta,\theta_1),m_k\in[1,2,\ldots,V] \tag{4}$$

2.2.2 角色预测 RP

除了通过全词掩码来学习语义知识之外，本文还考虑话语的角色信息(RP, Role Prediction)。需要注意的是，由于本文关注的双角色多轮对话，因此，角色预测属于二分类任务(这里用 A 和 C 来表示，A 表示客服代表，C 表示客户)。例如，对于对话 $d_k=\{(s_1,u_1),(s_2,u_2),\ldots,(s_m,u_m)\}$，将根据 u_1 对应对的嵌入向量，判断该片段是否由 s_1 表达的。图 1 案例的第二轮对话，预测其角色为 A；第七轮对话，预测其角色为 C。利用每个片段的第一个嵌入向量(即 E_{i1})，作为片段的语义表示，通过非线性角色分类器，预测该片段的角色，并与实际角色进行比较。

$$p_i^2=sigmoid(E_{i1}\cdot W_2+b_2) \tag{5}$$

其中，$p_i^2\in \boldsymbol{R}$，$W_2\in \boldsymbol{R}^{d\times 1}$，$b_2$ 是非线性分类器的偏置参数。然而，本文的模型不限于双角色对话，同样适合于多角色对话场景，只是由二分类任务，转换为多分类任务。非线性角色分类器的参数记为 θ_2，角色预测预训练任务的损失函数可以表示为

$$L_2(\theta,\theta_2)=-\sum_{r=1}^{m}logp_i^2(n=n_r\,|\,\theta,\theta_2),n_r\in[A,C] \tag{6}$$

2.2.3 话语内部反转预测 IQRP

多轮对话具有先天的内在顺序逻辑，因此，本文提出两种基于顺序的预训练任务。本部分从微观角度，先介绍角色的话语内部顺序(IQRP, Intra-Query Reverse Prediction)。角色在表达话语的过程中，往往包含多个句子，这些句子之间是天然具有先后顺序和内在逻辑的，如果将其中连续两个句子进行互换，那么，将影响句子的实际语义和含义。因此，QUERY 内部反转预测属于二分类任务，即存在反转和没有反转。例如，对于角色 s_i 的话语 $u_i=\{u_1^i,u_2^i,u_3^i,u_4^i,\ldots,u_n^i\}$，假定 u_1^i 与 u_2^i 组成一个句子，u_3^i 与 u_4^i 组成一个句子，如果选择该连续两个句子进行反转，即得到 $u_i=\{u_3^i,u_4^i,u_1^i,u_2^i,\ldots,u_n^i\}$。图 1 案例的第 10 轮对话的内部发生了反转，因此，预测为 True。利用每个片段的第一个嵌入向量，预测该片段内部是否存在句子反转，并与实际标签进行比较。

$$p_i^3=\mathrm{sigmoid}(E_{i1}\cdot W_3+b_3) \tag{7}$$

其中，$p_i^3\in \boldsymbol{R}$，$W_3\in \boldsymbol{R}^{d\times 1}$，$b_3$ 是非线性分类器的偏置参数。非线性反转分类器的参数记为 θ_3，QUERY 内部反转预测预训练任务的损失函数可以表示为

$$L_3(\theta,\theta_3)=-\sum_{t=1}^{m}logp_i^3(c=c_t\,|\,\theta,\theta_3),c_t\in[True,\ False] \tag{8}$$

2.2.4 轮次间互换预测 IQEP

从宏观角度来看，在一通多轮对话的过程中，角色是基于之前的多次交互信息，决定如何输出本轮次的话语，因此，不同轮次的话语也具有内在的顺序逻辑，它可能是对前一轮次的回复，也可能是对更早轮次的澄清或否定等。如果将任意两轮话语进行交换顺序，那么，必然将影响整个对话的实际语义和含义。因此，轮次间互换预测(IQEP, Inter-Query Exchange Prediction)，实际上是需要预测每一片段在整个对话的实际轮次，属于多分类任务。例如，对于对话 $d_k=\{(s_1,u_1),(s_2,u_2),(s_3,u_3),\ldots,(s_m,u_m)\}$，如果选择第 1 和 3 轮次对话互换，即得到 $d_k=\{(s_3,u_3),(s_2,u_2),(s_1,u_1),\ldots,(s_m,u_m)\}$，这里就需要根据片段 u_3 的嵌入向量，预测其轮次为 3(即使片段 u_3 在输入序列中处于第 1 轮次)，同理，根据片段 u_1 的嵌入向量，预测其轮次为 1(即使片段 u_1 在输入序列中处于第 3 轮次)。图 1 案例的第三轮和第四轮进行了互换，因此，需要预测输入第三轮实际为第四轮，而输入第四轮实际为第三轮。利用每个片段的第一个嵌入向量，预测该片段在对话中的轮次，并与实际轮次进行比较。

$$p_i^4=softmax(E_{i1}\cdot W_4+b_4) \tag{9}$$

其中，$p_i^4\in \boldsymbol{R}^S$，$W_4\in \boldsymbol{R}^{d\times S}$，$b_4$ 是非线性分类器的偏置参数。非线性轮次分类器的参数记为 θ_4，轮次预测预训练任务的损失函数可以表示为

$$L_4(\theta,\theta_4)=-\sum_{s=1}^{S}logp_i^4(e=e_s\,|\,\theta,\theta_4),e_s\in[1,2,\ldots,S] \tag{10}$$

2.3 多任务联合训练

最终，综合考虑上述四个自监督预训练任务，通过多任务联合学习，最小化上述损失函数之和，训练本文的 SPD-BERT 模型，模型总的损失函数可以表示为

$$\mathrm{L}(\theta,\theta_1,\theta_2,\theta_3,\theta_4)=L_1(\theta,\theta_1)+L_2(\theta,\theta_2)+L_3(\theta,\theta_3)+L_4(\theta,\theta_4) \tag{11}$$

2.4 模型预训练

结合上述四项自监督预训练任务，本部分详细介绍如何预训练 SPD-BERT 模型，包括：如何生成高质量、大规模训练语料，以及模型训练的参数设置。

2.4.1 语料数据

由于目前没有开源的大规模口语对话数据集，因此，本文收集金融领域内 2020 年 5 月至 2021 年 5 月的人工客服坐席的对话数据，这些数据是客户和客服代表使用口语通过电话方式进行对话，并将语音通过 ASR 转译后的文本数据。为了保护客户的隐私，这里将文本中出现的数字(包括但不限于：身份证号码、手机号码、金额、银行卡号、住址门牌号等)、姓名、地址等，全部进行随机替换脱敏处理。对于脱敏后的数据语料，为了提升语料质量，进行了如下预处理：1)针对常见的 ASR 转译错误(例如：备用金和被用金)，进行强制转换；2)为了满足双角色的基本要求，剔除只涉及 1 个角色的对话(例如：外呼未接听)；3)为了使得模型学习到更丰富的语义知识，并且，人人对话的总轮次往往较多，这里，剔除总轮次较少(8 轮及以下)的对话；4)剔除对话文本的总长度小于 80 的对话；5)由于本文基于 transformer 特征提取器，结合语料数据分析，对话文本的总长度限制在 486；6)考虑到轮次预测属于多分类任务，对话总轮次限制在 32 轮。经过预处理后，得到大约 2000 万通高质量、多轮次、双角色口语对话。

为了充分提高语料的利用率，本文采取动态生成训练样本的方式，具体体现在如下两个方面。一是，基于全量对话语料，构建领域专有词典，对于任意对话，随机选择其中若干专有术语，进行全词掩码。二是，对于三个自监督任务(RP、IQRP、IQEP)，并不需要对每个片段分别预测，而是随机选择其中部分片段进行预测。对于角色预测任务 RP，随机选择若干片段(而非全部片段)，预测其角色。同理，对于 IQRP 和 IQEP，也采取同样的处理方式。由于上述方法都基于随机选择，因此，对于任意一通对话，可以生成多个训练样本，从而大幅度增加训练样本的数量。特别是，生成的训练样本数

量与对话总轮次呈正相关。

另外，需要注意的是，考虑到片段的嵌入向量会用于 RP、IQRP、IQEP 任务，因此，应尽量避免同一个片段同时参与多个预测任务，而其他片段却没有参与到任务学习中。也就是说，随机选择若干个片段，预测其对应的角色。再从剩余的片段中，随机选择若干片段，预测其内部是否存在反转。然后，再从剩余的片段中，随机选择若干片段，预测其轮次。

2.4.2 模型训练

这里的 transformer 编码器配置与 BERT[5]的 $BERT_{base}$ 保持一致，并且，使用开源的中文 BERT 参数来初始化 transformer 编码器，学习率设置为 5e-5，使用学习率预热，非线性分类器的激活函数设置为 GELU[26]，优化器设置为 Adam[27]，批大小设置为 32，在 Tesla V100 上进行模型训练。对于所有实验，本文按照 80%、10%、10%的比例将数据集拆分成训练集、验证集、测试集，依据验证集的效果，选择最优模型，并在测试集进行评估。每组实验进行 4 次，取 4 次评估结果的平均值，作为最终的评估结果。

3　实验

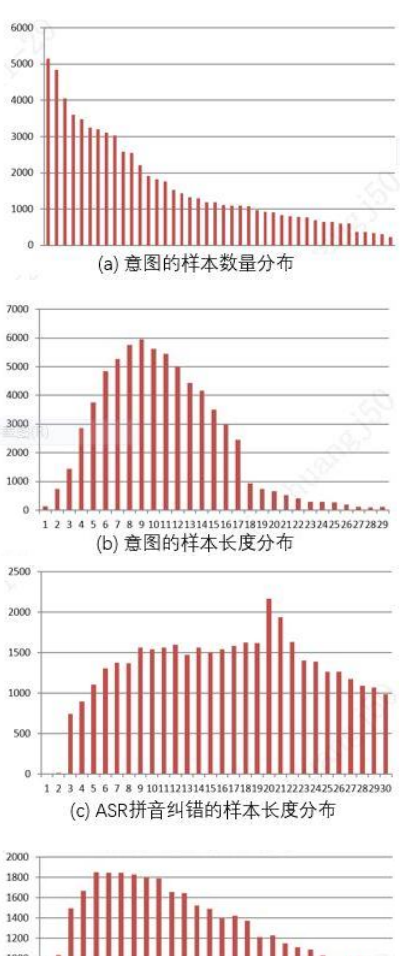
基于人人对话语料训练的 SPD-BERT 模型，是为了学习到口语对话的领域知识，可应用于下游的口语理解任务中，例如：智能质检，智能客服，智能助手等。这里，以智能客服场景为例，精调 SPD-BERT 模型，应用于三个典型口语语言理解下游任务：意图识别、ASR 拼音纠错和产品名识别，并比较不同模型在数据集上的效果。

3.1　实验数据

笔者所在机构在金融领域智能客服方面积累了大量意图识别训练数据，可以作为本次实验对象。另外，为了提升对话效果，笔者所在机构标注了相当数量的 ASR 拼音纠错训练数据和产品名(例如：理财产品，基金产品等)识别训练数据，三项任务的训练数据分布如图 3 所示。(a)(b)意图识别训练数据；(c)ASR 拼音纠错训练数据；(d)产品名识别训练数据。可以发现，在口语对话理解任务中，大部分样本是短文本，长度 32 个字符左右，这与常见的文档型数据存在较为明显的差异。表 1 展示了三个数据集的数据统计分析，可以发现，针对领域内特定任务，为了达到预期生产的效果，本文人工标注了大量的标签数据。值得注意的是，这些训练数据都是基于单轮对话的客户话语，因此，按照 SPD-BERT 模型的输入格式要求，在客户话语的首部添加[CLS]标识，并在尾部添加[SEP]标识，输入到模型中，得到每个位置的嵌入向量。对意图识别任务，提取[CLS]对应的嵌入向量，再增加全连接层，输出每个意图的得分和概率。例如，客户话语：帮我看看我卡里还有多少钱，对应的意图为：查余额。对于 ASR 拼音纠错任务，提取每个字符位置对应的嵌入向量，再增加全连接层，输出该字符是否存在转译错误(属于二分类任务)；对于存在转译错误的字符，使用另外一个全连接层，输出词典中每个字符的得分和概率(属于多分类任务)。例如，客户话语：我想数回马上到期的理财产品，其中，“赎回”被 ASR 错误转译为“数回”，这将严重影响后续的意图识别。对于产品名识别任务，提取每个字符位置对应的嵌入向量，再增加 CRF 层，输出每个字符属于产品名的得分和概率。例如，客户话语“光伏 50ETF 这款基金怎么样？”，输出基金产品名：光伏 50ETF。

3.2　实验结果

对于不同的下游任务，对比多种基线模型在各自测试数据集上的效果。对于意图识别模型，使用的对比模型包括：基于 CNN[28]的文本分类模型，基于 RNN 的文本分类模型，基于 BERT[5]的文本分类模型，基于 ERNIE[29]的文本分类模型。表 2 展示了意图识别任务的实验结果。结果表明，预训练语言模型 BERT 更能理解话语的语义信息，相对于 CNN、RNN 分别显著提升了 2.83%、1.56%，而 ERNIE 由于包含了更丰富的知识，F1 进一步提升了 0.47%，达到了非常好的效果。而本文的 SPD-BERT 模型考虑了角色信息、语义信息和对话结构信息，因此，相对于 ERNIE 进一步提升了 0.38%。该实验表明，SPD-BERT 在短文本分类任务上，具有明显的优势。



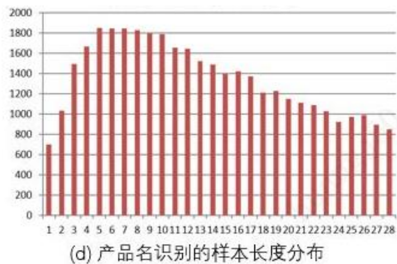


图 3 三个典型口语语言理解下游任务训练数据分析

Fig. 3　Statistics of three typical SLU tasks

表 1　数据集统计分析

Tab. 1　Statistics of train data, validation data and test data

数据集描述	意图识别数据集	ASR 拼音纠错数据集	产品名识别数据集
样本总数	66818	54928	37805
样本平均长度	12.64	18.61	17.24
意图数量	108	-	-
意图平均样本数量	618	-	-
训练集样本数量	53454	43942	30244
验证集样本数量	6681	5492	3780
测试集样本数量	6683	5494	3781

对于 ASR 拼音纠错任务，首先对每个字符进行错误检测，如果认为该字符存在拼音错误，则尝试纠正错误。选择包含了拼音和笔画的预训练模型 PLOME[16]作为本实验的基线模型。如表 3 所示，比较了 PLOME 和 SPD-BERT 的字符级别、句子级别的 P(精准率)、R(召回率)、F1 值(P 和 R 的调和平均数)。对于字符级别，在错误检测阶段，F1 显著提升了 1.1%；在错误纠正阶段，F1 也提升了 0.4%。对于句子级别，

在错误检测阶段，F1 提升较为明显，达到 2.8%；在错误纠正阶段，F1 也取得明显效果，提升了 1.8%。该实验表明，SPD-BERT 在 ASR 拼音纠错方面，具有明显的优势。

表 2　意图识别任务的实验结果

Tab. 2　Comparisons of experiment results of intent detection

模型	精准率(P)	召回率(R)	F1
TextCNN	0.9342	0.9337	0.9339
TextRNN	0.9443	0.9490	0.9466
BERT	0.9616	0.9629	0.9622
ERNIE	0.9665	0.9674	0.9669
SPD-BERT	**0.9701**	**0.9714**	**0.9707**

表 3　ASR 拼音纠错任务的实验结果

Tab. 3　Comparison of experiment results of ASR correction

模型	错误检测(字符)			错误纠正(字符)			错误检测(句子)			错误纠正(句子)		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
PLOME	0.914	0.784	0.844	0.869	0.681	0.764	0.707	0.661	0.683	0.623	0.583	0.602
SPD-BERT	**0.933**	**0.789**	**0.855**	**0.870**	**0.686**	**0.768**	**0.727**	**0.695**	**0.711**	**0.634**	**0.607**	**0.620**

对于产品名识别任务，选择常规的 BiLSTM+CRF[30]和 BERT+CRF 作为基线模型。表 4 展示了基于实体级别的实验结果，可以发现，由于理财产品覆盖数百款、基金产品覆盖数千款，而人工标注的数据量较为丰富，因此，名称识别的效果整体较好。BERT+CRF 相较于 BiLSTM+CRF，F1 提升了 1.87%，SPD-BERT+CRF 在 BERT 的基础上则进一步提升了 0.48%。该实验表明，SPD-BERT 在命名实体识别方面，也具有明显的优势。

表 4　产品名识别任务的实验结果

Tab. 4　Comparison of experiment results of product NER

模型	精准率	召回率	F1	准确率
BiLSTM+CRF	0.9427	0.9457	0.9442	0.9378
BERT+CRF	0.9634	0.9624	0.9629	0.9598
SPD-BERT+CRF	**0.9676**	**0.9679**	**0.9677**	**0.9646**

4　结束语

本文提出面向口语对话的预训练语言模型 SPD-BERT，并构建大规模人人口语对话语料。根据笔者的经验，该模型是首个面向口语对话、多轮次、双角色的语言模型。通过四个自监督预训练任务：全词掩码，角色预测，话语内部反转预测和轮次间互换预测，该模型不仅考虑话语的角色信息，还融合多轮对话结构和语义信息。通过在金融领域智能客服场景的三个典型下游任务中的详细实验，证明了该模型的有效性。

另外，本文的提出的第四个预训练任务：轮次间互换预测，仅仅考虑对话中任意两句 QUERY 的互换。可以考虑，基于 QUERY 对的轮次互换，也就是：$d_k=\{(s_1,n_1),(s_2,n_2),(s_3,n_3),(s_4,n_4),(s_5,n_5)\}$，可以转换为：$d_k=\{(s_3,u_3),(s_4,u_4),(s_1,u_1),(s_2,u_2),(s_5,u_5)\}$，模型需要同时预测 4 个位置的正确顺序。大量相关研究表明，基于 PAIR 的损失函数，比基于 ITEM 的往往带来性能上的提升。

在此基础上，笔者希望从如下三个方面，继续提升该模型的能力。一方面，继续扩大领域口语对话语料库，更大规模的语料，往往能够带来模型效果的提升。另外，尝试更加复杂的自监督预训练任务，学习到更复杂的语义、结构等信息，从而提升模型的能力。最后，探索基于该模型，应用于对话场景的其他任务，例如：高频意图识别，对话树自动构建，知识图谱构建，商机发现，个性化智能对话等等。

参考文献：

[1] 曹亚如，张丽萍，赵乐乐．多轮任务型对话系统研究进展 [J]．计算机应用研究，2021，39 (2)．(Cao Yaru, Zhang Liping, Zhao Lele. Research progress of multi-turn task-oriented dialogue system [J]. Application Research of Computers, 2021, 39 (2)

[2] 赵阳洋，王振宇，王佩，等．任务型对话系统研究综述 [J]．计算机学报，2020, 43 (10): 1862-1896. (Zhao Yangyang, Wang Zhenyu, Wang Pei, *et al.* A survey on Task-Oriented Dialogue Systems [J]. Chinese Journal of Computers, 2020, 43 (10): 1862-1896)

[3] 陈晨，朱晴晴，严睿，等．基于深度学习的开放领域对话系统研究综述 [J]．计算机学报，2019, 42 (7): 1439-1466. (Chen Chen, Zhu Qinqin, Yan Rui, *et al.* , Survey on Deep Learning Based Open Domain Dialogue System [J]. Chinese Journal of Computers, 2019, 42 (7): 1439-1466.)

[4] 杨宁，卢菁，邵清，等．基于无向分块加权图的无模式实体识别方法研究 [J]．计算机应用研究，2021, 38 (1): 169-174. (Yang Ning, Lu Jing, Shao Qing, *et al.* , Research on schema-agnostic entity resolution based on undirected block weighted graph [J]. Application Research of Computers, 2021, 38 (1): 169-174.)

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, *et al.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C]// North American Chapter of the Association for Computational Linguistics. 2019.

[6] Suman V Ravuri, Andreas Stolcke. Recurrent neural network and LSTM models for lexical utterance classification [C]// Conference of the International Speech Communication Association. 2015.

[7] Gakuto Kurata, Bing Xiang, Bowen Zhou, *et al.* Leveraging Sentence-level Information with Encoder LSTM for Semantic Slot Filling [C]// Empirical Methods in Natural Language Processing. 2016.

[8] Haihong E, Niu Peiqing, Zhongfu Chen, *et al.* A Novel Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling [C]// Meeting of the Association for Computational Linguistics. 2019.

[9] Libo Qin, Tailu Liu, Wanxiang Che, *et al.* A Co-Interactive Transformer for Joint Slot Filling and Intent Detection [C]// International Conference on Acoustics, Speech, and Signal Processing. 2021.

[10] Qian Chen, Zhu Zhuo, Wen Wang. BERT for Joint Intent Classification and Slot Filling [J]. arXiv: Computation and Language. 1902. 10909, 2019.

[11] Giuseppe Castellucci, Valentina Bellomaria, Andrea Favalli, *et al.* Multi-lingual Intent Detection and Slot Filling in a Joint BERT-based Model [J]. arXiv: Computation and Language. 1907. 02884, 2019.

[12] Libo Qin, Tianbao Xie, Wanxiang Che, *et al.* A Survey on Spoken Language Understanding: Recent Advances and New Frontiers [J]. arXiv: Computation and Language. 2103. 03095, 2021.

[13] Junjie Yu, Zhenghua Li. Chinese Spelling Error Detection and Correction Based on Language Model, Pronunciation, and Shape [C]// In Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 220-223, 2014.

[14] Yuzhong Hong, Xianguo Yu, Neng He, *et al.* FASPell: A Fast, Adaptable, Simple, Powerful Chinese Spell Checker Based On DAE-Decoder Paradigm [C]// Empirical Methods in Natural Language Processing. pages 160-169, 2019.

[15] Xingyi Cheng, Weidi Xu, Kunlong Chen, *et al.* SpellGCN: Incorporating Phonological and Visual Similarities into Language Models for Chinese Spelling Check [C]// In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 871-881, 2020.

[16] Shulin Liu, Tao Yang, Tianchi Yue, *et al.* PLOME: Pre-training with Misspelled Knowledge for Chinese Spelling Correction [C]// In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 2991-3000, 2021.

[17] Tomas Mikolov, Kai Chen, Greg S Corrado, *et al.* Efficient Estimation

of Word Representations in Vector Space [J]. arXiv: Computation and Language, 2013.

[18] 孙飞, 郭嘉丰, 兰艳艳, 等. 分布式单词表示综述 [J]. 计算机学报, 2019, 42 (7): 1605-1624. (Sun Fei, Guo Jiafeng, Lan Yanyan, *et al.* A Survey on Distributed Word Representation [J]. Chinese Journal of Computers, 2019, 42 (7): 1605-1624.) Peters *et al.* , Deep contextualized word representations. NAACL-HLT, 2018.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, *et al.* Attention is All you Need [C]// Neural Information Processing Systems. 2017.

[20] Alec Radford, Karthik Narasimhan, *et al.* , Improving language understanding by generative pre-training [C] , 2018, 1-12.

[21] Chien-Sheng Wu, Steven C H Hoi, *et al.* TOD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogue [C]// Empirical Methods in Natural Language Processing. 2020.

[22] Minjeong Kim, Gyuwan Kim, Sang Woo Lee, *et al.* ST-BERT: Cross-modal Language Model Pre-training For End-to-end Spoken Language Understanding [J]. arXiv: Computation and Language. 2021.

[23] Jia-Chen Gu, Chongyang Tao, Zhen-Hua Ling, *et al.* MPC-BERT: A Pre-Trained Language Model for Multi-Party Conversation Understanding [C]// Meeting of the Association for Computational Linguistics. 2021.

[24] 王孟宇, 俞鼎耀, 严睿, 等. 基于HRED模型的中文多轮对话任务方法研究 [J]. 中文信息学报, 2020, 34 (8): 78-85. (Wang Mengyu, Yu Dingyao, Yan Rui, *et al.* Chinese Multi-turn Dialogue Tasks based on HRED model [J]. Journal of Chinese Information Processing, 2020, 34 (8): 78-85.)

[25] Yiming Cui, Wanxiang Che, Ting Liu, *et al.* Pre-Training with Whole Word Masking for Chinese BERT [J]. arXiv: Computation and Language. arXiv: 1906. 08101, 2019.

[26] Dan Hendrycks, Kevin Gimpel. Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units [J]. arXiv: Learning. abs/1606. 08415, 2016.

[27] Diederik P Kingma, Jimmy Ba. Adam: A Method for Stochastic Optimization [C]// International Conference on Learning Representations. ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

[28] Yoon Kim. Convolutional Neural Networks for Sentence Classification [C]// Empirical Methods in Natural Language Processing. 2014.

[29] Zhengyan Zhang, Xu Han, Zhiyuan Liu, *et al.* ERNIE: Enhanced Language Representation with Informative Entities [C]// Meeting of the Association for Computational Linguistics. 2019.

[30] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, *et al.* Neural Architectures for Named Entity Recognition [C]// North American Chapter of the Association for Computational Linguistics. 2016.